

STAT 465: SURVIVAL ANALYSIS

Nana Kena Frempong (Ph.D.)

Dept. of Statistics and Actuarial Science
Kwame Nkrumah University of Science and Technology
Kumasi, Ghana

January 13, 2022

Content

- 1 Overview of Survival Analysis and Concepts
- 2 The Mathematics of Survival Models
- 3 Estimation of survival models
- 4 Inference of Survival Models
- 5 Intro to Cox-PH regression models
- 6 References

What is Survival Analysis?

Survival Analysis is the modern name given to the collection of statistical procedures which accommodate time-to-event censored data.

It typically focuses on **time to event** data. In the most general sense, survival analysis consists of techniques for positive valued random variables, such as

- time to death
- time to onset (or relapse) of a disease
- length of stay in a hospital
- duration of a strike
- money paid by health insurance
- viral load measurements
- time to finishing a doctoral dissertation!

Kinds of survival studies include:

- Clinical designs/Longitudinal
- Prospective cohort studies
- Retrospective cohort studies
- Cross-sectional studies

Typically, survival data are not fully observed, but rather are censored.

Cross-Sectional Studies

In this study design, we begin by defining the study group, an identifiable group of persons whose survival pattern is to be studied. Examples of the study group would be the entire population of a city, state or nation (demographic), or the policyholders of a life insurance company or members of a pension plan.

Longitudinal Studies

In contrast to the large sample cross-sectional studies, clinical studies are frequently quite different in their design. Rather than selecting an interval of time, and observing failure data within that observation interval, these studies select a study group and follow the experience of that group longitudinally into the future. This study design is feasible when the time until failure is, in general, fairly short.

The statistical tools treated are applicable to data from medical clinical trials, public health, epidemiology, engineering, economics, psychology, actuarial and demography as well.

Some knowledge of discrete data methods will be useful, since analysis of the “time to event” uses information from the discrete (i.e. binary) outcome of whether the event occurred or not.

Definitions and Notations

Failure time random variables are always *non-negative*. That is, if we denote the failure time by T , then $T \geq 0$.

T can either be **discrete** (taking a finite set of values, e.g. $a_1, a_2, \dots, a_n$) or **continuous** (defined on $(0, \infty)$).

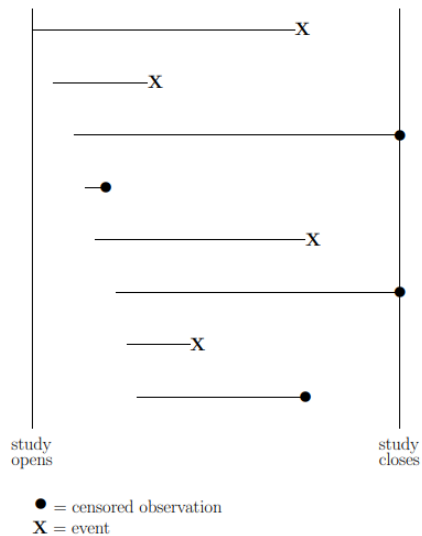
A random variable X is called a **failure time random variable** if $X = \min(T, U)$, where U is a nonnegative censoring variable.

Definitions and Notations

In order to define a failure time random variable, we need:

- an unambiguous **time origin** (e.g. randomization to clinical trial, purchase of car)
- a **scale** (e.g. real time (days, years), mileage of a car)
- definition of the **event** (e.g. death, relapse of a disease, withdrawal of life insurance policy, etc.)

Figure 1: Illustration of survival data



The illustration of survival data on the previous page shows several features which are typically encountered in analysis of survival data:

- individuals do not all enter the study at the same time
- when the study ends, some individuals still haven't had the event yet
- other individuals drop out or get lost in the middle of the study, and all we know about them is the last time they were still "free" of the event

The first feature is referred to as "**staggered entry**"

The last two features relate to "**censoring**" of the failure time events.

Censoring

Types of Censoring

- **Right-Censoring:** The r.v, $X_i = \min(T_i, U_i)$ is observed due to loss to follow-up, drop-out and study termination. We call this right-censored because the true unobserved event is to the right of our censoring time; i.e. all we know about is that the event has not happened at the end of follow-up or study.

E.g. For a plan study on mortality for two years, at the end of the study some life's were still alive. Even though we know life's may die in near. In addition to observing X_i , we also get to see the **failure indicator**:

$$\delta_i = \begin{cases} 1 & T_i \leq U_i \\ 0 & T_i > U_i \end{cases}$$

Censoring

- **Left-censoring:** can only observe $Y_i = \max(T_i, U_i)$. Define the failure indicators? E.g. (Miller) study of the age at which children learn a task. Some already knew (left-censored), some learned during the study (exact), others had not yet learned by the end of the study (right-censored).
- **Interval-censoring:** We observe (L_i, R_i) where $T_i \in (L_i, R_i)$.
**Right censoring is the common assumption we will deal with in survival analysis.*

Independent vs Informative censoring

- We say censoring is **independent** (non-informative) if U_i is independent of T_i .
 - E.g (1). If U_i is the planned end of the study (say, 2 years after the study opens), then it is usually independent of the event times.
 - E.g (2). If U_i is the time that a patient drops out of a study because he/she got much sicker and/or had to discontinue taking the study treatment, then U_i and T_i are probably not independent.
 - An individual censored at u_i should be representative of all subjects who survive to U_i .
- Censoring is considered **informative** if the distribution of U_i contains any information about the parameters characterizing the distribution of T_i

Exercise

- A** How does prospective cohort studies differ from retrospective studies?
- B** At KNUST, members of the Tertiary Education Workers Union (TEWU) have been on an indefinite strike since week one of reopening, pending stakeholder consultations to address their several concerns. Researchers from the university's Human Resource Department have attempted to investigate how long it will take workers to return to post within three weeks, supposing that members could do so with or without the approval of TEWU leadership. At the end of the study, some workers
- (i) had permanently resigned and taken up new work placements
 - (ii) are ill and responding to treatment at various health facilities
 - (iii) are waiting for their concerns to be addressed by leadership
 - (iv) had resumed work at their various departments

- 1 In this investigation, determine the
 - (a) event of interest
 - (b) time origin
 - (c) time scale
 - (d) time window
- 2 For each of the groups in (i) to (iv), discuss whether or not they have been censored.

The Mathematics of Survival Models

There are several equivalent ways to characterize the probability distribution of a survival random variable.

We will focus on the following terms:

- The probability density function $f(t)$ – *PDF*
- The cumulative distribution function $F(t)$ – *CDF*
- The survival distribution function $S(t)$ – *SDF*
- The hazard rate function $\lambda(t)$ – *HRF*
- The cumulative hazard function $\Lambda(t)$ – *CHF*

The Mathematics of Survival Models

The survival distribution function is given as:

$$S(t) = P(T \geq t)$$

A non-increasing function, $S(0) = 1$ and $S(\infty) = 0$.

Density function (or Probability Mass Function) for discrete r.v.'s Suppose that T takes values $a_1, a_2, \dots, a_k$

$$f(t) = Pr(T = t)$$

$$= \begin{cases} f_j & t = a_j, j = 1, 2, \dots, k \\ 0 & t \neq a_j, j = 1, 2, \dots, k \end{cases}$$

Survival distribution function for a discrete r.v T

$$\begin{aligned}
 S(t) &= \sum_{u \geq t} f(u) \\
 &= \sum_{a_i \geq t} f(a_i) \\
 &= \sum_{a_i \geq t} f_j
 \end{aligned}$$

The Hazard rate function for a discrete r.v. T

$$\begin{aligned}
 \lambda(a_j) \equiv \lambda_j &= Pr(T = a_j | T \geq a_j) \\
 &= \frac{Pr(T = a_j)}{Pr(T \geq a_j)} \\
 &= \frac{f(a_j)}{S(a_j)}
 \end{aligned}$$

The Mathematics of Survival Models

Probability density function for a continuous T:

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} Pr(t \leq T \leq t + \Delta t)$$

Survival distribution function for a continuous T:

$$S(t) = \int_t^{\infty} f(u) du$$

The Mathematics of Survival Models

Hazard Rate Function for a continuous r.v:

$$\begin{aligned}
 \lambda(t) &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \Pr(t \leq T \leq t + \Delta t | T \geq t) \\
 &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \frac{\Pr[(t \leq T \leq t + \Delta t) \cap (T \geq t)]}{\Pr(T \geq t)} \\
 &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \frac{\Pr(t \leq T \leq t + \Delta t)}{\Pr(T \geq t)} \\
 &= \frac{f(t)}{S(t)}
 \end{aligned}$$

In words, the rate of occurrence of the event at duration t equals the density of events at t , divided by the probability of surviving to that duration without experiencing the event.

1 Relationship between SDF and CDF!

$$S(t) = 1 - F(t)$$

2 Prove: Relationship between SDF and PDF, use results (1)

$$f(t) = -S'(t) \text{ or } S'(t) = -f(t)$$

3 Use relationship in (2) to prove,

$$\lambda(t) = -\frac{d}{dt}[\log S(t)]$$

4 Use relationship in (3) to prove,

$$\Lambda(t) = -\log S(t) + \log S(0)$$

$$S(t) = e^{-\Lambda(t)}$$

The Mathematics of Survival Models

Examples

- 1 Given that T follows the Uniform distribution over the interval $[0, \omega]$. Derive the survival and hazard functions of T .
- 2 Given that T follows the Exponential distribution with parameter μ . Derive the hazard, cumulative hazard and survival functions of T .
- 3 Given that T follows the Weibull distribution with parameters α and β . Given $S(t) = e^{-\alpha t^\beta}$, find the hazard and pdf of T

Soln

1

$$f(t) = \begin{cases} \frac{1}{\omega} & 0 < t < \omega \\ 0 & \text{elsewhere} \end{cases}$$

$$S(t) = P(T \geq t) = \int_t^{\omega} \frac{1}{u} du$$

$$= \frac{\omega - t}{\omega} \quad \omega > 0$$

$$h(t) = \frac{f(t)}{S(t)}$$

$$= \frac{1}{\omega - t} \quad 0 < t < \omega$$

2

$$f(t) = \mu e^{-\mu t} t \geq 0$$

$$S(t) = 1 - F(t)$$

$$= 1 - \int_0^t \mu e^{-\mu w} dw$$

$$= e^{-\mu t}$$

$$h(t) = \frac{\mu e^{-\mu t}}{e^{-\mu t}}$$

$$= \mu$$

$$H(t) = -\ln[S(t)]$$

$$= \mu t$$

Try (3)

The Mathematics of Survival Models

Exercise

A Derive the following, where T is continuous

- (i) The First Moment of the Random variable T ?
 $E(T) = \dots$
- (ii) The Second Moment of the Random Variable T ?
 $E(T^2) = \dots$
- (iii) The Variance of T ?
 $Var(T) = \dots$
- (iv) The Median of T ?

The Mathematics of Survival Models

Exercise

- B** The failure time T , is modelled by the *log-logistics* distribution with *pdf*

$$f(t) = \begin{cases} \frac{\alpha \lambda t^{\alpha-1}}{(1 + \lambda t^\alpha)^2} & t \geq 0, \lambda > 0, \alpha > 0 \\ 0 & \text{otherwise} \end{cases}$$

Determine

- (i) the survival function, $S(t)$, and calculate $S(3)$ given $\lambda = 1$ and $\alpha = 2$.
- (ii) the hazard function, $h(t)$, and describe the form of the function
- (iii) an expression for the median survival time, t_m , in terms of α and λ .

Estimation of Survival and Hazard Functions

- We can estimate the survival (or hazard) function in two ways:
 - By developing an empirical estimate of the survival function(i.e., non-parametric estimation)
 - By specifying a parametric model for $\lambda(t)$ based on a particular density function $f(t)$.
- **If no censoring:** The empirical estimate of the survival function, $\tilde{S}(t)$, is the proportion of individuals with event times greater than t .
- **With censoring:** if there are censored observations, then $\tilde{S}(t)$ is not a good estimate of the true $S(t)$.

Estimation of Survival and Hazard Functions

- Non-parametric methods must be used to account for censoring (life-table methods, Kaplan-Meier estimator, e.t.c)
- Here we discuss the most famous non-parametric approach for estimating the survival distribution, called the **Kaplan-Meier Estimator**.
- To motivate the derivation of this estimator, we will first consider a set of survival times where there is no censoring.

Estimation of Survival and Hazard Functions

Example 1

The following are times to relapse (weeks) for 21 leukemia patients receiving control treatment (Table 1.1 of Cox & Oakes):

1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23

How would we estimate $S(10)$, the probability that an individual survives to 10 weeks or later? What about $S(8)$?

We make use of the empirical estimate

$$\tilde{S}(t) = \frac{\# \text{individuals with } T \geq t}{\text{Total sample size}}$$

Soln:

Values of t	$\tilde{S}(t)$
$t \leq 1$	$^{21}_{/21} = 1.000$
$1 < t \leq 2$	$^{19}_{/21} = 0.905$
$2 < t \leq 3$	$^{17}_{/21} = 0.809$
$3 < t \leq 4$	$^{16}_{/21} = 0.762$
$4 < t \leq 5$	$^{14}_{/21} = 0.667$
$5 < t \leq 8$	$^{12}_{/21} = 0.571$
$8 < t \leq 11$	$^8_{/21} = 0.381$
$11 < t \leq 12$	$^6_{/21} = 0.286$
$12 < t \leq 15$	$^4_{/21} = 0.190$
$15 < t \leq 17$	$^3_{/21} = 0.143$
$17 < t \leq 22$	$^2_{/21} = 0.095$
$22 < t \leq 23$	$^1_{/21} = 0.048$
$t > 23$	$^0_{/21} = 0.000$

From the table above: $\tilde{S}(10) = 0.381$ and $\tilde{S}(8) = 0.571$

Example 2

The following times show the number of hours it took 20 rats who ate poison to die. Estimate the survival and hazard functions.

1,1,1,1,1,1,1,1,3,3,3,3,3,4,4,5,5,10,10,12

Soln:

a_j	$P(T = a_j)$	$\tilde{S}(t)$	$\tilde{\lambda}(t)$
1	0.4	1.00	0.40
3	0.25	0.60	0.42
4	0.10	0.35	0.29
5	0.10	0.25	0.40
10	0.10	0.15	0.67
12	0.05	0.05	1.00

Estimation of Survival and Hazard Functions

- The Example 1 shows the empirical survival function when there is no censoring. *What if there is censoring?*
- There is one oldest estimator in literature called Life table (*Actuarial Estimator*) which will not be considered here.
- The Kaplan-Meier (or KM) estimator is probably the most popular approach. It can be justified from several perspectives:
 - Product limit estimator
 - Likelihood justification
 - Redistribute to the right estimator
- We will start with an intuitive motivation based on conditional probabilities, and then review some of the other justifications.

Conditional Probability

In a 1958 paper in the Journal of the American Statistical Association, Kaplan-Meier proposed a way to non-parametrically estimate $S(t)$, even in the presence of censoring. The method is based on the ideas of **conditional probability**.

A quick review of conditional probability:

- *Conditional Probability*: Suppose A and B are two events. Then,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- *Multiplication law of probability*: this can be obtained from the above relationship, by multiplying both sides by $P(B)$:
We get $P(A \cap B) = P(A|B)P(B)$

Extension to two or more finite events:

Suppose $A_1, A_2, \dots, A_k$ are k different events. Then, the probability of all k events happening together can be written as a product of conditional probabilities:

$$\begin{aligned} P(A_1, A_2, \dots, A_k) &= P(A_1 \cap A_2 \dots \cap A_k) \\ &= P(A_k | A_{k-1} \cap A_{k-2} \cap \dots \cap A_1) \times \\ &\quad P(A_{k-1} | A_{k-2} \cap \dots \cap A_1) \times \dots \times P(A_2 | A_1) \times P(A_1) \end{aligned}$$

Now lets apply these ideas to estimate $S(t)$ Suppose $a_k < t \leq a_{k+1}$, then

$$\begin{aligned}
 S(t) &= P(T \geq a_{k+1}) \\
 &= P(T \geq a_1, T \geq a_2, \dots, T \geq a_{k+1}) \\
 &= P(T \geq a_1) \times \prod_{j=1}^k P(T \geq a_{j+1} | T \geq a_j) \\
 &= \prod_{j=1}^k [1 - P(T = a_j | T \geq a_j)] \\
 &= \prod_{j=1}^k [1 - \lambda_j]
 \end{aligned}$$

So,

$$\hat{S}(t) \cong \prod_{j=1}^k \left(1 - \frac{d_j}{r_j}\right) = \prod_{j:a_j \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

- Where d_j is the number of events at a_j and r_j is the number at risk at a_j .
- The KM Estimator of the survivorship function is given as:

$$\hat{S}_{KM}(t) = \prod_{j:\tau_j \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

Where

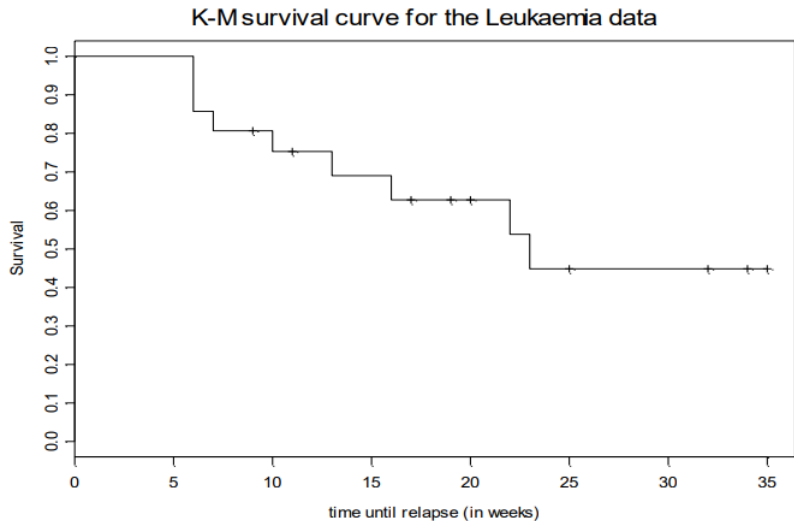
- $\tau_1, \tau_2, \dots, \tau_k$ is the set of k distinct event times observed in the sample.
- d_j is the number of events at τ_j .
- r_j is the number of individuals “at risk” right before the j^{th} event time (individuals experiencing the event or censored at or after that time).
- c_j is the number of censored observations between the j^{th} and $(j + 1)^{th}$ event times. Censoring tied at event times are included in c_j .
- Note: two useful formulas are:
 - $r_j = r_{j-1} - d_{j-1} - c_{j-1}$
 - $r_j = \sum_{l \geq j} (c_l + d_l)$

Example 3

We consider the treated group of Leukemia from Table 1.1 of Cox and Oakes: $6^+, 6, 6, 6, 7, 9^+, 10^+, 10, 11^+, 13, 16, 17^+, 19^+, 20^+, 22, 23, 25^+, 32^+, 32^+, 34^+, 34^+, 35^+$ [Note: times with + are right censored] Compute $S(6)$ and $S(7)$?

Soln:

t_j	d_j	r_j	$1 - \frac{d_j}{r_j}$	$\hat{S}_{KM}(t)$
0	0	21	1	1
6	3	21	$18/21$	$18/21$
7	1	17	$16/17$	$96/119$
10	1	15	$14/15$	$64/85$
13	1	12	$11/12$	$176/255$
16	1	11	$10/11$	$32/51$
22	1	7	$6/7$	$64/119$
23	1	5	$6/7$	$160/357$



Note that:

$\hat{S}(t^+)$ only changes at death(failure) times

$\hat{S}(t^+)$ is 1 up to the first death time

$\hat{S}(t^+)$ only goes to zero if the last event is death

Other Justifications for KM Estimator

Likelihood- based derivation (Cox and Oakes) For a discrete failure time variable, define:

- d_j is the number of failures at t_j .
- r_j the number of individuals “at risk” at t_j .
- λ_j Pr(death) in j^{th} interval (conditional on survival to start of interval)
- The likelihood is that g independent binomials:

Properties of the KM Estimator

Where n is the number of individuals in the study. This is just like an estimated probability from a binomial distribution so we have:

$$\hat{S}(t) \sim N\left(S(t), \frac{S(t)[1-S(t)]}{n}\right)$$

From Large sample theory

- $\hat{S}(t)$ is still approximately normal
- The mean converges to the true mean $S(t)$
- The variance is a bit more complicated since the denominator n includes some censored observations
- Once we get the variance then we can construct (pointwise) confidence intervals

Methods for computing variance of the KM estimator:

Delta Method

If Y is a random variable that is normal with mean μ and variance σ^2 , then the $g(Y)$ is approximately normally distributed with mean $g(\mu)$ and variance $[g'(\mu)]^2 \sigma^2$

We make use of two specific examples:

A $Z = \log(Y)$

Then $Z \sim N \left[\log(\mu), \left(\frac{1}{\mu}\right)^2 \sigma^2 \right]$

B $Z = \exp(Y)$

then $Z \sim N [e^\mu, (e^\mu)^2 \sigma^2]$

* $\hat{\lambda}_j$ is approximately normal with mean λ_j and variance

$$\frac{\hat{\lambda}_j(1-\hat{\lambda}_j)}{r_j}$$

Greenwoods Formulae

$$\begin{aligned}
 \text{var}(\log[\hat{S}(t)]) &= \sum_{j:\tau_j < t} \text{var}(\log(1 - \hat{\lambda}_j)) \\
 &= \sum_{j:\tau_j < t} \left(\frac{1}{1 - \hat{\lambda}_j} \right)^2 \text{var}(\hat{\lambda}_j) \\
 &= \sum_{j:\tau_j < t} \left(\frac{1}{1 - \hat{\lambda}_j} \right)^2 \hat{\lambda}_j(1 - \hat{\lambda}_j)/r_j \\
 &= \sum_{j:\tau_j < t} \frac{\hat{\lambda}_j}{(1 - \hat{\lambda}_j)r_j} \\
 &= \sum_{j:\tau_j < t} \frac{d_j}{(r_j - d_j)r_j}
 \end{aligned}$$

From;

$$\begin{aligned}\hat{S}(t) &= \exp[\log(\hat{S}(t))] \\ \text{var}(\hat{S}(t)) &= [\hat{S}(t)]^2 \text{var}(\log(\hat{S}(t)))\end{aligned}$$

We obtain the Greenwood formula:

$$\text{var}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j: \tau_j < t} \frac{d_j}{(r_j - d_j)r_j}$$

Confidence Intervals

- For a 95% confidence interval, we could use, $\hat{S}(t) \pm 1.96se[\hat{S}(t)]$ formula. **Problem: this approach can yield values > 1 or < 0**
- **Better approach:** Get a 95% confidence interval for $L(t) = \log(S(t))$ since this quantity is restricted over $(-\infty, 0)$, the confidence interval will be in proper range when we transform back.
- Lets see why this works:

$$0 \leq \hat{S}(t) \leq 1$$

$$-\infty \leq \log[\hat{S}(t)] \leq 0$$

Summary of Confidence Intervals on $S(t)$

- Calculate $\hat{S}(t) \pm Z_{1-\frac{\alpha}{2}} se[\hat{S}(t)]$, where $se(.)$ is calculated using Greenwood's formula is, and replace negative lower bounds by 0 and upper bounds greater than 1 by 1. Results not very satisfactory, recommended by *Collett*.
- Use a log transformation to stabilize the variance and allow for non-symmetric confidence intervals.

$$\log(\hat{S}(t)) \pm Z_{1-\frac{\alpha}{2}} se[\log(\hat{S}(t))]$$

Nelson-Aalen Estimator

Estimating the survivorship function using the Cumulative Hazard (Nelson-Aalen Estimator)

- Suppose we want to estimate $\Lambda(t) = \int_0^t \lambda(u) du$.
- Think about dividing the observed time span in a series of fine intervals, so that there is only one event per interval:
- We can approx. the CH(Cumulative Hazard) as :

$$\hat{\Lambda}(t) = \sum_j \lambda_j \Delta$$

- We further approximate

$$\hat{\Lambda}(t) = \sum_j \lambda_j \Delta = \sum_{j: \tau_j \leq t} \frac{d_j}{r_j}$$

- We know that the CH changes only at event times
- Hence we write Nelson-Aalen Estimator as

$$\sum_{j:\tau_j \leq t} \frac{d_j}{r_j}$$

Fleming-Harrington Estimator

Once we have the NA estimator we can also find another estimator of $S(t)$ called **Fleming-Harrington**:

$$\hat{S}_{FH}(t) = \exp(-\hat{\Lambda}_{NA}(t))$$

NB:

In general as the sample size increases, this estimator of the survival function will be close to the Kaplan-Meier estimator

Table 2: Illustrating the Nelson-Aalen estimator

		D	D		C	C	D
r_j	n	n	n-1	n-2	n-2	n-3	n-4
d_j	0	1	1	0	0	0	1
c_j	0	0	0	0	1	1	0
$\hat{\lambda}(t_j)$	0	$\frac{1}{n}$	$\frac{1}{n-1}$	0	0	0	$\frac{1}{n-4}$
$\hat{\Lambda}(t_j)$	0	$\frac{1}{n}$	$\frac{1}{n} + \frac{1}{n-1}$	$\frac{1}{n} + \frac{1}{n-1}$	$\frac{1}{n} + \frac{1}{n-1}$	$\frac{1}{n} + \frac{1}{n-1}$	$\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-4}$

Table 3: Example from Cox-Oaks data

From the Cox-Oaks data, the Nelson Aalen and Fleming Herrington estimators are computed as

t_j	d_j	r_j	$\hat{\Lambda}_{NA}(t)$	$\hat{S}_{FH}(t)$
0	0	21	0	1
6	3	21	$1/7$	0.8669
7	1	17	$24/119$	0.8174
10	1	15	$479/1785$	0.7646
13	1	12	$837/2380$	0.7035
16	1	11	0.4426	0.6424
22	1	7	0.5855	0.5569
23	1	6	0.7521	0.4714

Exercise

The population of elderly people in a prison is observed during the period 1st January 1994 to 31st December 1996. The duration of residence (measured to the nearest number of months) is recorded for those who die during the period, for those who are released from the prison during the period and for those who are still in residence on 31st December 1996.

The recorded data measured in months are:

6⁺, 6, 6, 6, 7, 9⁺, 10⁺, 10, 11⁺, 13, 16, 17⁺, 20, 23⁺

where + indicates those who were released from the prison during the period or who were still in residence on 31 December 1996.

- (a) Calculate the product-limit (Kaplan-Meier) and Flemming Herrington estimates of the survival function, $S(t)$, where t is the duration of residence in the prison.
- (b) Calculate the median survival time.

Comparison of Survival Curves

- We have been looking at some nonparametric approaches for estimating the survival function overtime for a single sample of individuals.
- Should we base our overall comparison of on:
 - The further distance between the two curves?
 - The median survival for each group?
 - A rank-based test? That is, we could rank all of the event times and then see whether the sum of ranks for one group was less than the other.

Comparison of Survival Curves

All of these are pretty reasonable options,
There have been several proposals for how to compare the survival of two groups.

For the moment, we are sticking to nonparametric comparisons

Why nonparametric?

- Fairly robust
- Efficient relative to parametric tests
- Often simple and intuitive

Comparison of Survival Curves

Logrank Test: Formal Definition

- The logrank test is obtained by constructing a (2×2) table at each distinct death time, and comparing the death rates between the two groups, conditional on the number at risk in the groups.
- It does not matter which group you choose to sum over.
- The logrank statistic depends on ranks of event times only.
- If the estimated survival curves crosses, then there is evidence that the hazards are not **proportional**.

Logrank Test statistics

Formal Definition:

$$\chi^2_{logrank} = \frac{[\sum_{j=1}^k (d_{0j} - r_{0j}d_j/r_j)]^2}{\sum_{j=1}^k \frac{r_{1j}r_{0j}d_j(r_j - d_j)}{r_j^2(r_j - 1)}}$$

$$\chi^2_{logrank} = \frac{\left(\sum_j (o_j - e_j)\right)^2}{\sum_j v_j}$$

Where :

$$o_j = d_{0j}$$

$$e_j = r_{0j}d_j/r_j$$

$$v_j = r_{1j}r_{0j}d_j(r_j - d_j)/[r_j^2(r_j - 1)]$$

The test statistic is approximately Chi-square with $Df = 1$

Example 4

Two groups of patients were followed up on their survival status after treatment. The survival data (in weeks) are:

Group A: 14, 34⁺, 25, 46, 56⁺, 30

Group B: 25⁺, 34, 43⁺, 46, 34, 36

- 1 Compute the logrank test statistic
- 2 What is your conclusion regarding the comparison of these two groups in terms of survival?

Soln

Ordered time	d_{0j}	r_{0j}	d_{1j}	r_{1j}	d_j	r_j	e_j	$o_j - e_j$	v_j
14	1	6	0	6	1	12	0.500	0.500	0.250
25	1	5	0	6	1	11	0.455	0.545	0.248
30	1	4	0	5	1	9	0.444	0.556	0.247
34	0	3	2	5	2	8	0.750	-0.750	0.402
36	0	2	1	3	1	5	0.400	-0.400	0.240
46	1	2	1	1	2	3	1.333	-0.333	0.222

$$o_j = d_{0j}, e_j = r_{0j}d_j/r_j, v_j = r_{1j}r_{0j}d_j(r_j - d_j)/[r_j^2(r_j - 1)]$$

$$\chi^2_{logrank} = \frac{\left(\sum_j (o_j - e_j)\right)^2}{\sum_j v_j} = \frac{(0.118)^2}{1.609} = 0.00865$$

Exercise

Two groups of patients are followed up on their survival status after treatment. The survival data (in weeks) are

Group A : 9⁺, 5, 7, 3, 18

Group B : 20, 20⁺, 12, 19, 33⁺

Compute the log-rank statistics. What is your conclusion regarding the comparison of these two groups in terms of their survival at 5% level of significance? [Hint: $\chi^2_{0.025,1} = 5.024$]

Cox-PH Model

General Framework for Survival Analysis

We observe (X_i, δ_i, Z_i) for individual i , where

- X_i is a censored failure time random variable
- δ_i is the failure/censored indicator
- Z_i represents a set of covariates

Cox-PH Model

Modeling of Survival data

- We will explore the relationship between survival and explanatory variables.
- Z_i represents a set of covariates:
 - Continuous factors(age, blood pressure...)
 - Discrete factors(gender, marital status,...)
 - Possible interactions
 - A discrete covariate A with a-levels

Cox-PH Model

Cox-PH Model:

$$\lambda(t; Z) = \lambda_0(t) \exp(\beta Z)$$

- This is an example of a **semi-parametric** model
- For two treatments, we define the covariate $Z = 1$ or $Z = 0$
 $\lambda_1(t) = \lambda(t; z = 1) = \lambda_0(t)e^{\beta z} = \lambda_0(t)e^{\beta}$
 $\lambda_0(t) = \lambda(t; z = 0) = \lambda_0(t)e^{\beta z} = \lambda_0(t)$
- Why proportional hazard?
 The hazard of the two groups remains proportional overtime.
- What is the hazard ratio?
- What is the baseline hazard function?

Cox-PH Model

The Cox-PH model is a linear model for the log of the hazard ratio

Estimation of model parameters

- The Partial Likelihood method
- We define the following
 $R(X_i)$ - risk set of the failure time for individual i
 $R(\tau_j)$ - risk set at the j^{th} failure time

$$L_i(\beta) = S(X_i) = \exp \left[- \int_0^{X_i} \lambda_i(u) du \right]$$

for individual censored at X_i

$$L_i(\beta) = S(X_i) \lambda_i(X_i) = \lambda_i(X_i) \exp \left[- \int_0^{X_i} \lambda_i(u) du \right]$$

for individual fails at X_i

Total likelihood contributions

$$L(\beta) = \prod_{i=1}^n \lambda_i(X_i)^{\delta_i} \exp \left[- \int_0^{X_i} \lambda_i(u) du \right]$$

We multiply and divide the total likelihood by $\left[\sum_{j \in R(X_i)} \lambda_i(X_i) \right]^{\delta_i}$

$$\begin{aligned} L(\beta) &= \prod_{i=1}^n \left[\frac{\lambda_i(X_i)}{\sum_{j \in R(X_i)} \lambda_i(X_i)} \right]^{\delta_i} \\ &= \prod_{i=1}^n \left[\frac{\lambda_0(X_i) \exp(\beta Z_i)}{\sum_{j \in R(X_i)} \lambda_0(X_i) \exp(\beta Z_j)} \right]^{\delta_i} \\ &= \prod_{i=1}^n \left[\frac{\exp(\beta Z_i)}{\sum_{j \in R(X_i)} \exp(\beta Z_j)} \right]^{\delta_i} \end{aligned}$$

Cox-PH Model

Partial Likelihood

Remarks on the partial likelihood:

- Contributions only at the death times
- The partial likelihood is NOT a product of independent terms, but of conditional probabilities

Cox-PH Model

Example 5

Individual	X_i	δ_i	Z_i
1	9	1	4
2	8	0	5
3	6	1	7
4	10	1	3

Find the partial likelihood contributions at each failure time.

Cox-PH Model

Soln: Ordered data

Individual	X_i	δ_i	Z_i
3	6	1	7
2	8	0	5
1	9	1	4
4	10	1	3

$$\ell_3(\beta) = \frac{e^{7\beta}}{e^{7\beta} + e^{5\beta} + e^{4\beta} + e^{3\beta}}, \ell_2(\beta) = 1, \ell_1(\beta) = \frac{e^{4\beta}}{e^{4\beta} + e^{3\beta}}, \ell_4(\beta) = \frac{e^{3\beta}}{e^{3\beta}}$$

$$\text{Where; } \ell_i(\beta) = \left[\frac{\exp(\beta Z_i)}{\sum_{j \in R(X_i)} \exp(\beta Z_j)} \right]^{\delta_i}$$

The partial likelihood would be the product of these four terms.

Cox-PH Model

Model Fitting

$$\beta = (\beta_1, \beta_2, \dots, \beta_p)$$

$$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$$

$$\text{solving } u(\hat{\beta}) = 0$$

$$u(\beta) = \left(\frac{\partial \log L(\beta)}{\partial \beta_1}, \frac{\partial \log L(\beta)}{\partial \beta_2}, \dots, \frac{\partial \log L(\beta)}{\partial \beta_p} \right)$$

Example 6

An investigation has been carried out on sickness inception rates for a particular class of business that a company sells. The following proportional hazards regression was fitted to the data:

$$h_i(t) = h_0(t) \exp [0.02(x_i - 45) - 0.01(45 - y_i) + 0.04s_i + 0.03o_i]$$

Where:

- $x_i = \text{Max}(\text{age of the } i^{\text{th}} \text{ policyholder, } 45)$
- $y_i = \text{Min}(\text{age of the } i^{\text{th}} \text{ policyholder, } 45)$
- $s_i = 1$ if the i^{th} life is a smoker
- 0 if the i^{th} life is a non-smoker
- $o_i = 0$ if the i^{th} life's job is clerical
- 1 if the i^{th} life's job is light manual
- 2 if the i^{th} life's job is heavy manual

Questions

- Define the type of covariate represented by each of the values of x_i , y_i , s_i and o_i .
- Describe the class of lives to which the baseline hazard function applies.
- What does the model (if correct) tell you about the survival function of a clerical worker compared to that of a heavy manual worker if all other covariates are identical?

Exercise

You are investigating the survival times of patients who have just undergone heart surgery at one of 3 hospitals – **A**, **B**, or **C**. You have recorded the following data for each patient.

$$Z_1 = \begin{cases} 0 & \text{for females} \\ 1 & \text{for males} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{if patient attended Hospital B} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_3 = \begin{cases} 1 & \text{if patient attended Hospital C} \\ 0 & \text{otherwise} \end{cases}$$

Exercise conti...

You have decided to model the force of mortality at time t (measured in days since the operation was performed) by an equation of the form $\lambda(t) = \lambda_0(t)e^{BZ^T}$, and you have estimated the parameter values to be:

$$\hat{B}_1 = 0.031 \quad \hat{B}_2 = -0.025 \quad \hat{B}_3 = 0.011$$

Compare the force of mortality for a ***female*** patient who attended Hospital **A** with that of

- (a) a ***female*** patient who attended Hospital **B**
- (b) a ***male*** patient who attended Hospital **C**

References

Some useful references

- *Collett: Modelling Survival Data in Medical Research*
- *Cox and Oakes: Analysis of Survival Data*
- *Kalbfleisch and Prentice: The Statistical Analysis of Failure Time Data*
- *Lee: Statistical Methods for Survival Data Analysis*
- *Fleming & Harrington: Counting Processes and Survival Analysis*
- *Hosmer & Lemeshow: Applied Survival Analysis*
- *Kleinbaum: Survival Analysis: A self-learning text*